scientific reports

Check for updates

OPEN

Risk sensitive twin distributional critics with a lambda lower confidence bound for continuous control reinforcement learning

Onur Osman[1], Bahar Yalcin Kavus[2], Tolga Kudret Karaca[3]✉, Gokalp Tulum[1] & Mehmet Ali Karabulut[3]

Off-policy actor–critic methods such as Twin Delayed Deep Deterministic Policy Gradient (TD3) are the workhorse of continuous-control reinforcement learning. However, they rely on scalar value estimates and offer no explicit way to control risk in temporal-difference targets. We introduce Twin Distributional Critics with λ-Lower Confidence Bound (TDC-λ), a TD3-style algorithm that learns two distributional critics and, for each transition, forms its target from a lower confidence bound of the form (μ − λσ) across critics. The risk parameter λ smoothly interpolates between a distributional TD3 limit and increasingly conservative targets. A single implementation supports either a deterministic actor or a tanh-squashed Gaussian policy, while evaluation always uses the deterministic mean action. We evaluate TDC-λ on five standard MuJoCo benchmarks HalfCheetah-v4, Hopper-v4, Ant-v4, Walker2d-v4, and Humanoid-v4 against strong entropy-regularized baselines. Across tasks, TDC-λ matches or improves final return while consistently reducing variance. Sweeping λ further shows that stronger penalties on high-variance critics improve robustness on challenging, high-dimensional domains. These results indicate that distributional critics combined with simple risk-sensitive target selection can substantially improve stability in off-policy reinforcement learning without sacrificing sample efficiency.

Keywords Reinforcement learning, Continuous control, Distributional reinforcement learning, Risk-sensitive control, Actor–critic methods

Continuous-control reinforcement learning (RL) has gathered around off-policy actor–critic methods that learn value estimates and policies from replayed experience while controlling approximation bias. Lillicrap et al.[1] introduced Deep Deterministic Policy Gradient (DDPG), demonstrating that deterministic policy gradients paired with replay and target networks could scale to high-dimensional action spaces. Fujimoto et al.[2] proposed Twin Delayed Deep Deterministic Policy Gradient (TD3), mitigating overestimation with clipped double critics, target-policy smoothing, and delayed actor updates, which together deliver robust gains on MuJoCo benchmarks[2]. These ideas draw on Van Hasselt's Double Q-learning and its deep variant by Van Hasselt et al.[3], which showed that decoupling action selection from evaluation reduces positive bias in temporal-difference targets; temporal-difference learning itself traces back to Sutton's seminal formulation[3,4].

In parallel, maximum-entropy (MaxEnt) RL has formalized stochastic control by augmenting return with policy entropy. Haarnoja et al.[5] introduced Soft Actor–Critic (SAC), an off-policy actor–critic that optimizes a stochastic policy under an entropy-regularized objective and has become a standard for robust exploration in continuous control[5]. Earlier, Haarnoja et al.[6] proposed deep energy-based policies, casting the critic as an energy function and motivating soft-value training that later informed modern MaxEnt algorithms. Despite their empirical success, conventional MaxEnt methods typically alternate *policy evaluation* and *policy improvement* and approximate the soft value with Monte-Carlo estimates, which can introduce variance and optimization mismatch between actor and critic[5,6]. To address these issues, Chao et al.[7] introduced Maximum Entropy Reinforcement Learning via Energy-Based Normalizing Flow (MEOW), formulation that unifies evaluation and improvement into a single-objective training process, provides an *exact* soft-value expression.

[1]Department of Electric Electronics Engineering, İstanbul Topkapi University, 34087 Istanbul, Turkey. [2]Quality Coordination Office, İzmir Katip Çelebi University, 35620 Izmir, Turkey. [3]Department of Industrial Engineering, Istanbul Topkapi University, 34087 Istanbul, Turkey. ✉email: tolgakudretkaraca@topkapi.edu.tr

Natively models multi-modal action distributions with efficient sampling, while empirically observing that deterministic inference can outperform stochastic sampling at test time in several continuous-control tasks[7]. Flow architectures such as Real NVP[8], Glow[9], Masked Autoregressive Flow[10] and Neural Spline Flows[11] supply the invertible transformations that make EBFlow practical in high dimensions[8,11].

Recent TD3-lineage extensions refine stability and data efficiency by rethinking decision aggregation, trust-region control, and asynchronous updates. Nachum et al.[12] proposed Trust-PCL to stabilize off-policy updates with a path-consistency trust region; Gu et al.[13] introduced Q-Prop to fuse policy gradients with an off-policy critic; Meng et al.[14] developed an off-policy TRPO variant with a monotonic improvement guarantee; and Wu et al.[15] presented A-TD3 to accelerate convergence via adaptive asynchronous update[12,15]. Domain-specific TD3 variants further expand the design space by coupling entropy-maximizing exploration[16], multi-critic aggregation for complex driving scenes[17], and human-in-the-loop advice in continuous action spaces[16,18]. Within this trend, Osman et al.[19] introduced AdvB-TD3, which augments TD3 with a cooperative advisory board scored by a shared critic and dynamic member management; the authors reported faster convergence, higher returns, and reduced variability across MuJoCo tasks such as BipedalWalker-v3, HalfCheetah-v4, and Humanoid-v4, thereby underscoring the value of structured selection on top of a strong TD3 backbone[19]. This paper positions Twin Distributional Critics with λ-Lower-Confidence Bound (TDC-λ) as a compact, risk-aware extension of TD3 that integrates distributional value learning with conservative target selection while preserving the off-policy, sample-efficient training loop. When λ = 0, TDC-λ reduces to a distributional TD3 variant with a clipped-double-style selection: in that case the LCB score $\mu - \lambda\sigma$ collapses to the mean and simply chooses the critic with the smaller mean; larger λ values yield increasingly conservative targets. Specifically, we propose twin quantile critics that predict return distributions and a λ-tuned aggregation that selects, for each transition, the critic with the better mean–variance score ($\mu -$ λσ), thereby reducing estimation drift under noisy targets without sacrificing TD3's stabilizers[2]. Complementing the deterministic actor used by default as is customary for TD3-style exploitation the same training pipeline admits a stochastic actor head, in this work, a tanh-squashed Gaussian policy, enabling entropy-regularized exploration in regimes where robustness to multi-modality or partial observability matters[5,11]. The same interface could also host flow-based policies[7]. This bridge directly reflects EBFlow's insight that one can reap the benefits of stochastic training while deploying deterministically when it performs better in practice[7]. Compared to AdvB-TD3's[19] critic-guided action selection ensemble, TDC-λ operates at the target-formation level by shaping distributional critics and risk sensitivity; the two approaches are complementary within the TD3 ecosystem and point to a unifying theme of bias-aware, selection-aware control.

Our main contributions are threefold; in summary, TDC-λ draws on the determinism-for-exploitation strengths of TD3[2], the variance-control logic behind Double Q-learning and temporal-difference training[3,4] and the expressivity and exact-value advantages of modern MaxEnt formulations[5,11]. The result is a single framework that (i) learns distributional critics, (ii) performs risk-aware target selection via λ, and (iii) supports deterministic or stochastic policies under one off-policy pipeline, aligning with contemporary evidence on when each inference mode is most effective.

Methodology

This section formalizes our algorithmic design, defines the learning objectives, and details the training loop used in our implementation. We build on the stabilizing principles behind TD3 twin critics, policy delay, and target-policy smoothing while replacing scalar critics with distributional critics and introducing a risk-sensitive target selection governed by a non-negative parameter λ. Here, λ denotes the lower-confidence bound (LCB) risk parameter in the critic selection term $\mu_i - \lambda\sigma_i$. We further provide a binary switch ζ∈{0,1} that instantiates either a deterministic actor μ_θ or a stochastic actor π_θ (a|s) with a tanh-Gaussian sampling distribution. When ζ = 1, the actor is trained under a standard MaxEnt objective with a learnable temperature α, but the critics still approximate the non-soft return distribution using the same λ-LCB target. At evaluation time, we always deploy the deterministic mean action, in line with prior observations that deterministic inference can outperform stochastic sampling in continuous-control MaxEnt RL.

Problem setting and notation

We consider an infinite-horizon discounted MDP with continuous state space $\mathcal{S}$, continuous action space $\mathcal{A}$, transition density p_T, reward $R : \mathcal{S} \times \mathcal{A} \to R$ and discount $\gamma \in (0, 1)$. At time t, the agent observes $s_t \in \mathcal{S}$, chooses $a_t \in \mathcal{A}$, receives $r_t = R(s_t, a_t)$ and transitions to $s_{t+1} \sim p_T(\cdot s_t, a_t)$. We employ an experience replay buffer $\mathcal{D}$ and a mini-batch size B.

TDC −λ maintains two distributional action-value estimators ("critics") $Z_{\varphi_1}, Z_{\varphi_2}$ that approximate the return distribution via N_q quantiles $\left\{Z^k_{\varphi_i}(s,a)\right\}_{k=1}^{N_q}$ at fixed locations $\tau_k = (2k-1)/2N_q$. The policy is parameterized by θ and can be run in two modes. Throughout, we use $\varphi = (\varphi_i, \varphi_i)$ for critic parameters and θ for actor parameters:

- Deterministic head $\mu_\theta : \mathcal{S} \to \mathcal{A}, (\zeta = 0)$
- Stochastic head $\pi_\theta(as)$ with a *tanh-Gaussian* sampling distribution $(\zeta = 1)$.

Target networks $\varphi'_1, \varphi'_2, \theta'$ are maintained by Polyak averaging with factor $\tau \in (0, 1)$. We denote by $\alpha \geq 0$ the entropy temperature (used only if $\zeta = 1$), by $\sigma_{\text{tgt}} > 0$ the target-smoothing noise scale, and by $\bar{c} > 0$ the clipping bound for target noise. In the stochastic configuration (ζ = 1) we parameterize the temperature as log α and optimize it online to match a target policy entropy H target = − dim(), following the standard automatic entropy tuning used in SAC[5].

Risk-sensitive target via λ-lower-confidence bound (λ-LCB)

For each transition (s, a, r, s', d) in a mini-batch, we generate a target action $\tilde{a}$ using the delayed target actor $\overline{\theta}$.

Deterministic case ($\zeta = 0$). We follow TD3 and set Eq. 1

$$\tilde{a} = clip\left(\mu_{\overline{\theta}}\left(s'\right) + \varepsilon, -a_{max}, a_{max}\right), \varepsilon \sim clip\left(\mathcal{N}\left(0, \sigma_{targ}^2 I\right), -c, c\right) \tag{1}$$

followed by clipping $\tilde{a}$ to the valid action range.

Stochastic case ($\zeta = 1$). We instead draw

$$\widetilde{a} \sim \pi_{\overline{\theta}}\left(\cdot|s'\right) \tag{2}$$

from the tanh-Gaussian target policy and do not inject additional TD3-style smoothing noise.

We evaluate both target critics distributional at $(s', \tilde{a})$ to obtain quantile sets

$$Z_i\left(s', \tilde{a}\right) = \left\{, z_{i,j}\left(s', \tilde{a}\right),\right\}_{j=1}^{N_q}, i \in \{1, 2\}. \tag{3}$$

For each critic i, we compute the distributional mean and dispersion

$$\mu_i\left(s', \tilde{a}\right) = \frac{1}{N_q}\sum_{j=1}^{N_q} z_{i,j}\left(s', \tilde{a}\right), \quad \sigma_i\left(s', \tilde{a}\right) = \sqrt{\frac{1}{N_q}\sum_{j=1}^{N_q}\left(z_{i,j}\left(s', \tilde{a}\right) - m_i\left(s', \tilde{a}\right)\right)^2} \tag{4}$$

and select the index

$$i^*\left(s', \tilde{a}\right) = \arg\min_{i \in \{1,2\}}\left(\mu_i\left(s', \tilde{a}\right) - \lambda\,\sigma_i\left(s', \tilde{a}\right)\right) \tag{5}$$

which recovers TD3's clipped-double selection when $\lambda = 0$ and becomes increasingly conservative as λ grows. The chosen critic defines the target quantile vector.

Let Z denote the return random variable represented by a distributional critic for a given transition $(s', \tilde{a})$ and let $\mu = \mathbb{E}[Z]$ and $\sigma = Std(Z)$ be estimated from the critic's quantile outputs. The score $\mu - \lambda\sigma$ admits a principled interpretation as a one-sided lower confidence bound. In particular, the one-sided Chebyshev-Cantelli inequality implies the distribution-free bound; $Pr\{Z \geq \mu - \lambda\sigma\} \geq \frac{\lambda^2}{1+\lambda^2}\lambda \geq 0$. So $\mu - \lambda\sigma$ is a conservative lower bound whose implied confidence increases monotonically with λ. Under an approximate Gaussian assumption, $\mu - \lambda$ also coincides with a "λ-sigma" lower quantile. We use this bound to select the more pessimistic of the twin target critics for target formation, thereby reducing the likelihood of propagating over-optimistic TD targets. This selection recovers TD3-style clipped double-Q behavior when λ = 0 (min over means) and becomes increasingly conservative as λ grows.

The parameter λ controls the conservativeness of the TD target selection through the lower-confidence-bound score $\mu - \lambda\sigma$ computed from each critic's predicted return distribution. Setting λ = 0 recovers a distributional analogue of TD3's clipped-double target selection (mean-only), while larger λ increasingly favors critics with narrower predicted distributions and thus more conservative targets. In practice we found that λ values in a compact range (e.g., λ ∈ [0,1]) are sufficient on MuJoCo tasks. As a rule of thumb, tasks that exhibit high critic dispersion and unstable learning (typically high-dimensional or long-horizon domains) tend to benefit from larger λ (more conservative targets), whereas tasks with consistently low critic dispersion can use smaller λ without sacrificing stability. When transferring TDC-λ to a new task, λ can be selected with a small-budget pilot sweep over a few candidate values and chosen by early validation return; stability diagnostics such as the dispersion of the selected target critic and/or minibatch TD-error variance can be used as additional signals to detect overly optimistic targets.

$$Z^*\left(s', \tilde{a}\right) = Z_{i^*}\left(s', \tilde{a}\right), \tag{6}$$

and we form per-quantile TD targets as

$$y_j = r + \gamma\,(1 - d)\; z_j^*\left(s', \tilde{a}\right), j = 1, \ldots, N_q. \tag{7}$$

Crucially, the target has the same form for $\zeta = 0$ and $\zeta = 1$; entropy regularization only enters through the actor objective described in Section II-D.

Quantile regression critics and loss

Each critic Z_{φ_i} outputs N_q quantiles at the fixed locations $\{\tau_k\}_{k=1}^{N_q}$. For every state–action in the mini-batch and every pair (k, j),

$$\delta_{k,j}^{(i)} = y_j - Z_{\varphi_i}^k(s, a). \tag{8}$$

We minimize the quantile-Huber loss

$$\rho_{\tau_k}(\delta) = |\tau_k - 1\{\delta < 0\}|\, L_\kappa(\delta), \tag{9}$$

where $L_\kappa(\cdot)$ is the Huber loss with threshold $\kappa > 0$. The critic objective is

$$L_Q = \frac{1}{|B|} \sum_{(s,a,\cdot)\in B} \sum_{i=1}^{2} \frac{1}{N_q^2} \sum_{k=1}^{N_q} \sum_{j=1}^{N_q} \rho_{\tau_k}\left(\delta_{k,j}^{(i)}\right). \tag{10}$$

Actor objectives and deterministic–stochastic toggle

We update the actor every d critic steps. In both modes, we define the scalar value estimate as the mean of the first critic's quantiles

$$Q_{\varphi_i}(s,a) = \frac{1}{N_q} \sum_{j=1}^{N_q} z_{1,j}(s,a). \tag{11}$$

Deterministic head ($\zeta = 0$). The deterministic actor $\mu_\theta(s)$ is trained to maximize this value:

$$J_{\det}(\theta) = E_{s\sim\mathbb{D}}\left[Q_{\varphi_1}(s, \mu_\theta(s))\right]. \tag{12}$$

Stochastic head ($\zeta = 1$). When the stochastic actor $\pi_\theta(as)$ is enabled, we use the standard reparameterized MaxEnt objective

$$J_{\text{stoch}}(\theta) = E_{s\sim D,\, a\sim\pi_\theta(\cdot s)}\left[\alpha \log \pi_\theta(as) - Q_{\varphi_1}(s,a)\right], \tag{13}$$

and update θ by descending $\nabla_t J_{\text{stoch}}(\theta)$. In practice, this expectation is implemented with a single tanh-Gaussian sample per state. This matches the soft-actor update in SAC, with the important difference that Q_{φ_1} is a non-soft distributional Q-function trained with the λ-LCB target.

In the $\zeta = 1$ configuration we also optimize the temperature via

$$L(\alpha) = E_{s,a\sim\pi_\theta}\left[\alpha\left(-\log \pi_\theta(as) - \mathcal{H}_{\text{target}}\right)\right] \tag{14}$$

where $\mathcal{H}_{\text{target}} = -\dim(\mathcal{A})$ is the target entropy. In practice we parameterize $\alpha = \exp(\log\alpha)$ and update $\log\alpha$ with a separate Adam optimizer. Regardless of whether we train with $\zeta = 0$ or $\zeta = 1$, evaluation uses the deterministic mean action

$$a_{\text{eval}}(s) = \tanh(\mu_\theta(s)) \tag{15}$$

which has been reported to outperform stochastic sampling at test time in several continuous-control MaxEnt settings.

Target policy smoothing, noise clipping, and Polyak averaging

In the deterministic configuration ($\zeta = 0$) we apply standard TD3-style target-policy smoothing: before feeding the target action into the critics, we add zero-mean Gaussian noise $\varepsilon \sim \mathcal{N}\left(0, \sigma_{targ}^2 I\right)$, clip it element-wise to $[-c, c]$, and finally clip the resulting action to the valid range $[-a_{max},\ a_{max}]$. When $\zeta = 1$ we do not add extra smoothing noise, since the tanh-Gaussian target policy is already stochastic; in that case $\tilde{a}$ is obtained directly from $\pi_{\overline{\theta}}(\cdots')$. Target parameters are updated by Polyak averaging,

$$\varphi_i' \leftarrow \tau\,\varphi_i + (1-\tau)\ \varphi_i',\ \ \theta_i' \leftarrow \tau\,\theta_i + (1-\tau)\ \theta_i' \tag{16}$$

and the actor is updated only once every d critic updates (policy delay). These stability heuristics mirror the empirically effective TD3 design. Moreover, $\zeta \in \{0,1\}$ toggles deterministic/stochastic training and $\lambda \geq 0$ controls risk sensitivity.

When $\zeta = 0$ and $\lambda = 0$, the update reduces to a TD3-style clipped-double target with a deterministic policy trained via distributional critics. When $\zeta = 1$, the critic's targets are unchanged, as in λ–LCB, while the actor minimizes the entropy-regularized objective J_{stoch} with a learnable temperature; evaluation uses the deterministic mean action.

Putting these components together, Algorithm 1 summarizes the complete TDC-λ training loop, including environment interaction, λ-LCB target construction, distributional critic updates, and the delayed deterministic/stochastic actor and temperature updates.

Algorithm 1: TDC-(λ)

Inputs:
Discount $\gamma \in (0,1)$;
risk $\lambda \geq 0$;
temperature α (learned only if ζ=1; learning rate η_α);
number of quantiles N_q with locations $\tau_k = \left(k - \frac{1}{2}\right)/N_q$;
Huber threshold κ;
target-smoothing std. σ_{targ} and clip c;
action bounds $[-a_{max}, a_{max}]$, Polyak factor τ;
policy delay d;
learning rates $\eta_Q, \eta_\pi, \eta_\alpha$;
entropy toggle $\zeta \in \{0$ (deterministic), 1 (stochastic)$\}$;
replay buffer $\mathcal{D}$; mini-batch size B;
Initialize:
Critics φ_1, φ_2 and targets $\varphi_1' \leftarrow \varphi_1$, $\varphi_2' \leftarrow \varphi_2$; actor θ and target $\theta' \leftarrow \theta$.
Deterministic head (ζ=0): $a = \mu_\theta(s)$.
Stochastic head (ζ=1): $\pi_\theta(\cdot \mid s)$ (tanh-Gaussian).
Even when ζ=1 (stochastic training), the environment uses the deterministic mean action;
$a_{\text{env(s)}} = \mu_\theta(s)$; sampling $a \sim \pi_\theta(\cdot \mid s)$ is used only inside the off-policy updates.
For *each environment step:* $t = 1,2,\ldots$ do
# Action (environment interaction)
Observe s_t. Draw exploration noise $\varepsilon_{\text{explore}} \sim \mathcal{N}(0, \sigma^2_{\text{explore}} I)$
Let $\mu_\theta(s_t)$ denote the deterministic mean action (either the TD3-style actor or the mean of the tanh-Gaussian head).
Set: $a_t = \begin{cases} \mu_\theta(s_t) & \text{if } \zeta = 1, \\ clip\left(\mu_\theta(s_t) + \varepsilon_{\text{explore}}, -a_{max}, a_{max}\right) & \text{if } \zeta = 0 \end{cases}$
Execute a_t, observe $r_t, s_{t+1}, d_t \in \{0,1\}$ and push $(s_t, a_t, r_t, s_{t+1}, d_t)$ into $\mathcal{D}$.

Sample a mini-batch and update
sample B transitions {(s, a, r, s′, d)} from $\mathcal{D}$, with |B| = B
#Target construction (λ–LCB, distributional)
For each $(s, a, r, s', d) \in B$,
$\left(\tilde{a}', \log \pi'\right) = \begin{cases} \left(\text{clip}\left(\mu_{\theta'}(s') + \varepsilon_{\text{tgt}}, a_{min}, a_{max}\right), 0\right), & \zeta = 0, \quad \varepsilon_{\text{tgt}} \sim \mathcal{N}\left(0, \sigma^2_{\text{tgt}}\right), \quad \left|\varepsilon_{\text{tgt}}\right| \leq \bar{c} \\ \left(\tilde{a}' \sim \pi_{\theta'}(\cdot \mid s'), \ \zeta = 1\right) \end{cases}$

Note that $\log \pi'$ is not used if $\zeta = 0$, on the other hand $(\tilde{a}, \log \pi') =$ sample from $\pi_{\theta'}(\cdot \mid s')$ record $\log \pi'$ for actor/temperature only.
Evaluate twin target critics at $\left(s', \tilde{a}'\right)$ to obtain quantiles $\{Z^j_{\varphi_1'}\}_{j=1}^{N_q}$, $\{Z^j_{\varphi_2'}\}_{j=1}^{N_q}$. Compute μ_i, σ_i and

$$i^\star = \arg\min_{i \in \{1,2\}} (\mu_i - \lambda \sigma_i)$$

For $j = 1, \ldots, N_q$,

$$\left(y_j = r + \gamma(1-d)\left[Z^j_{\varphi_{i^\star}}\left(s', \tilde{a}'\right)\right]\right)$$

#Critic update (quantile Huber).
For $i \in \{1,2\}$, $k = 1, \ldots, N_q$, $j = 1, \ldots, N_q$,

$$\delta^{(i)}_{k,j} = y_j - Z^k_{\varphi_i}(s,a), \quad \rho_{\tau_k}(\delta) = \left|\tau_k - 1\{\delta < 0\}\right| L_\kappa(\delta).$$

$$L_Q = \frac{1}{|B|} \sum_{(s,a,\cdot) \in B} \sum_{i=1}^{2} \frac{1}{N_q^2} \sum_{k=1}^{N_q} \sum_{j=1}^{N_q} \rho_{\tau_k}\left(\delta^{(i)}_{k,j}\right)$$

$$\left(\varphi_1 \leftarrow \varphi_1 - \eta_Q \nabla_{\varphi_1} L_Q, \quad \varphi_2 \leftarrow \varphi_2 - \eta_Q \nabla_{\varphi_2} L_Q\right)$$

#Delayed actor & target updates.
If $t \bmod d = 0$, update the actor:
Deterministik (ζ = 0):

$$J_{\text{det}}(\theta) = \frac{1}{|B|} \sum_{s \in B} Q_{\phi_1}\left(s, \mu_\theta(s)\right), \qquad L_{det}(\theta) = -J_{det}(\theta)$$

$$\theta \leftarrow \theta - \eta_\pi \nabla_\theta L_{det}$$

Stokastik (ζ = 1):

$$J_{\text{stoch}}(\theta) = \frac{1}{|B|} \sum_{s \in B} E_{a \sim \pi_\theta(\cdot \mid s)}\left[\alpha \log \pi_\theta(a \mid s) - Q_{\phi_1}(s,a)\right],$$

$$\theta \leftarrow \theta - \eta_\pi \nabla_\theta J_{stoch}(\theta)$$

Temperature (only if $\zeta = 1$):

$$\alpha = \exp(log\alpha),$$

$$L_\alpha(log\alpha) = -\frac{1}{|B|} \sum_{s \in B} E_{a \sim \pi_\theta}\left[\log \alpha \left(\log \pi_\theta(a \mid s) + \mathcal{H}_{\text{target}}\right)\right],$$

$$log\alpha \leftarrow log\alpha - \eta_\alpha \nabla_{log\alpha} L_\alpha,$$

$$\mathcal{H}_{\text{target}} = -\dim(\mathcal{A})$$

Polyak within the same delayed block:

$$\varphi_i' \leftarrow \tau\varphi_i + (1-\tau)\varphi_i' \ (i = 1,2), \qquad \theta' \leftarrow \tau\theta + (1-\tau)\theta'.$$

End For

When and $\lambda = 0$ the update reduces to a TD3-style clipped-double target with a deterministic policy trained using distributional critics: the λ-LCB score collapses to the plain mean, so $I^*(s', \tilde{a})$ selects the critic with the

smaller mean, while the critics are still optimized with the quantile-Huber loss. When $\zeta = 1$, the critic targets are unchanged and still follow the same λ-LCB construction; the soft term $-\alpha \backslash \log \pi_\theta\,(\mathrm{a}|\mathrm{s})$ does not enter the target. Instead, the actor minimizes the entropy-regularized objective

$$J_{\text{stoch}}\,(\theta) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_\theta(\cdot|s)}\left[\alpha \log \pi_\theta\,(a \mid s) - Q_{\varphi_1}\,(s,a)\right], \tag{17}$$

with a learnable temperature α, and evaluation uses the deterministic mean action $a_{eval}\,(s) = \mu_\theta\,(s)$.

Network parameterization and implementation

Each critic maps state–action pairs to N_q quantiles using an MLP backbone. The actor family comprises a deterministic mapping $f_\theta\,(s)$ and a tanh-Gaussian stochastic policy $\pi_\theta\,(a|s)$; in our implementation. we instantiate either the deterministic head or the stochastic head depending on ζ. We define the deterministic mean action as;

$$\mu_\theta\,(s) = tanh\,(f_\theta\,(s))\,, a_{max}, \tag{18}$$

which lies in $[-\mathrm{a}_{\max},\ \mathrm{a}_{\max}]^{\dim(\mathcal{A})}$. Replay uses uniform sampling. During data collection, we follow TD3 practice and add Gaussian exploration noise to this mean action; the environment executes

$$\varepsilon_{\text{explore}} \sim \mathcal{N}\left(0,\ \sigma^2_{\text{explore}} I\right) a_{env}\,(s) = clip\,(\mu_\theta\,(s) + \varepsilon_{explore}, -\mathrm{a}_{\max},\ \mathrm{a}_{\max}) \tag{19}$$

where clipping is applied element-wise to enforce the action bounds. When $\zeta = 1$, the stochastic head π_θ is used only inside the off-policy updates target sampling and actor/temperature gradients, while interaction with the environment still uses the deterministic mean $a_{env}\,(s)$ defined above. Deterministic actor (mapping) and Stochastic actor (tanh-Gaussian sampling) given in Eq. 20 and Eq. 21, respectively.

$$a = \mu_{\theta(s)} = \tanh\,(f_\theta\,(s)) * a_{\max.} \tag{20}$$

$$\mu_\theta\,(\mathrm{s})\,, \sigma_\theta\,(\mathrm{s}) > 0, \varepsilon \sim \mathcal{N}\,(0,\mathrm{I})\,, \widetilde{a} = \mu_\theta\,(s) + \sigma_\theta\,(s) \odot \varepsilon, a = \tanh\,(\widetilde{a}) * a_{\max.} \tag{21}$$

Deterministic versus stochastic instantiations

TDC λ can be instantiated with either actor:

(i) $\zeta = 0$ yields a deterministic TDC-λ variant that behaves like TD3, except that the critics are distributional and trained with the λ-LCB target;
(ii) $\zeta = 1$ enables a tanh-Gaussian policy $\pi_\theta\,(as)$ trained with the MaxEnt actor loss $J_{stoch}\,(\theta)$ and automatic temperature tuning. In both cases, interaction with the environment uses the deterministic mean action plus Gaussian exploration noise, while the stochastic head is only used inside the off-policy update when $\zeta = 1$. Unless stated otherwise, we report deterministic results and include the stochastic variant for completeness and robustness analysis.

Complexity and hyperparameters

Per gradient step, critic updates scale as $\mathcal{O}\left(2BN_q^2\right)$ due to the all-pairs quantile losses, while actor updates scale as $\mathcal{O}\,(B)$ for $\zeta = 0$ and as $\mathcal{O}\,(B)$ with a single reparameterized sample per state for $\zeta = 1$. The additional cost versus scalar critics is dominated by N_q^2; moderate N_q (e.g., 25–64) balances fidelity and speed. Typical settings include policy delay $d \in \{2, 3\}$, target noise $\sigma_{\text{tgt}} \in \{0.1, 0.3, 0.5\}$ with $\overline{c} \in \{0.1, 0.3, 0.5\}$, and Polyak factors $\tau \in \left\{5 \cdot 10^{-4}, 5 \cdot 10^{-3}\right\}$. The risk parameter λ is swept to assess sensitivity.

Experiments and results

This section evaluates TDC-λ on the five standard MuJoCo continuous-control tasks: HalfCheetah-v4**,** Hopper-v4**,** Ant-v4**,** Walker2d-v4**,** and Humanoid-v4. Unless stated otherwise, all curves are step-aligned (bin size 20 k) and report the mean performance with shaded variability across independent runs. We compare against *vanilla TD3, DDPG,* SAC[5] and MEOW[7], using our own re-implementations following their public descriptions. We train up to 1.5M steps on HalfCheetah and Hopper, 4.0M on Ant and Walker2d, and 5.0M on Humanoid. During training we periodically run deterministic evaluation for 5 episodes as in[7] and log the average return; evaluation code is shared across environments. Moreover, TDC-λ keeps twin distributional critics that output 64 quantiles and are trained with the quantile-Huber objective. The target distribution is formed from the lower-confidence bound (LCB) of the two critics, mean − λσ**,** computed per transition; the policy is then optimized against the selected critic. Target-policy smoothing (Gaussian noise with standard deviation σ and clipping c), a policy delay, and Polyak averaging together stabilize training. The actor can be deterministic (TD3-style) or stochastic (tanh-Gaussian) with automatic temperature". Shared architecture uses 256–256 MLPs for actor and critic. Furthermore, replay uses a 1 M-transition buffer and 256-batch size. Learning rates are 2e − 4–3e − 4 depending on the task, and τ is 0.0005–0.005 which is smaller on Humanoid. The full per-environment hyperparameter settings for TDC-λ are summarized in Table 1. Furthermore, we selected the risk parameter λ via an automated hyperparameter search using Ray Tune. For each environment, we evaluated a small discrete set [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] as the range of candidate λ values under a fixed compute budget and selected the configuration that maximized early validation performance. The final per-environment λ values are reported in

Environment	HalfCheetah-v4	Walker2d-v4	Humanoid-v4	Ant-v4	Hopper-v4
Total environment steps	1,500,000	4,000,000	5,000,000	4,000,000	1,500,000
Warm-up random steps	25,000	30,000	5000	5000	5000
Risk parameter (λ)	1	1	1	0.9	0.9
Discount (γ)	0.99	0.99	0.99	0.99	0.99
Target-policy noise σ/clip c	0.2/0.5	0.2/0.5	0.2/0.5	0.2/0.5	0.2/0.5
Policy delay (d)	2	2	2	2	2
Polyak factor (τ)	0.005	0.005	0.0005	0.005	0.005
Learning rate (actor/critic)	3e−4/3e−4	3e−4/3e−4	3e−4/2e−4	2e−4/2e−4	2e−4/2e−4
Number of Quantiles (Nq)	64	64	64	64	64
Quantile-Huber κ	1	1	1	1	1
Replay size/batch	1,000,000/256	1,000,000/256	1,000,000/256	1,000,000/256	1,000,000/256
Exploration noise	Action noise $\sigma = 0.1$	Action noise $\sigma = 0.1$	Action noise $\sigma = 0.1$	Action noise $\sigma = 0.1$	Action noise $\sigma = 0.1$
Architecture (actor/critic)	MLP 256–256/256–256; critic outputs 64 quantiles	Same	Same	Same (deterministic actor)	Same (deterministic actor)
Evaluation	Deterministic mean action	Deterministic mean action	Deterministic mean action	Noise-free action	Noise-free action
	Report 5-episode avg	Report 5-episode avg	Report 5-episode avg	Report 5-episode avg	Report 5-episode avg

Table 1. Final environment-specific hyperparameters for TDC-λ on the MuJoCo benchmark suite.

Table 1. We additionally provide a λ sensitivity analysis to illustrate how performance and stability vary with λ in Supplemantary material Fig. S1.

The scalar $\lambda \geq 0$ controls the conservatism of the target-selection rule through the score $\mu - \lambda\sigma$ computed from each critic's predicted return distribution. Setting $\lambda = 0$ reduces TDC-λ to a distributional TD3-style target selection based on the mean (a clipped-double analogue), whereas larger λ values increasingly penalize critics with high dispersion and thus bias targets toward more conservative estimates. In practice, λ can be treated as a user-facing risk-sensitivity knob: higher λ typically reduces target dispersion and stabilizes learning (at the cost of more conservative updates), while lower λ can speed up learning when critic dispersion is benign. As an actionable recipe, we recommend starting from a conservative default (e.g., $\lambda \approx 1$) in high-dimensional or noisy environments and decreasing λ when the critics' dispersion remains consistently low. For the experiments reported in this paper, λ was selected per environment using Ray Tune on a small training budget, optimizing deterministic evaluation return; the final results are then obtained by re-running the full training horizon with multiple random seeds using the selected λ. The resulting per-environment λ values are reported in Table 1, and Fig. 3 provides an explicit sensitivity sweep illustrating how performance changes with λ on representative tasks.

Additionally, SAC follows its standard stochastic-actor update with automatic entropy tuning. MEOW is reproduced from the EBFlow paper to the extent needed for algorithmic comparison on MuJoCo; recall that MEOW trains with a single objective and can be evaluated deterministically, an observation we also examine for TDC-λ. Full training/evaluation loops and environment settings used to produce the curves are provided in the per-task drivers and the agent modules cited above.

Figure 1 summarizes returns across all five MuJoCo tasks. TDC-λ achieves the highest or statistically comparable final performance on Ant-v4, Walker2d-v4, and HalfCheetah-v4, while also exhibiting visibly lower variance in the late stages of training. We attribute this stability to the risk-sensitive target that replaces optimistic targets with a distributional lower-confidence bound $\mu - \lambda\sigma$, computed per transition and used to form the quantile targets for both critics. This mechanism tempers over-estimation whenever critic dispersion is high and leads to steadier improvement as training progresses.

On the hardest benchmark, Humanoid-v4, TDC-λ achieves the strongest asymptotic performance surpassing TD3 and DDPG and matching or exceeding the remaining baselines despite the longer horizon and higher dimensionality; the smaller target-network update rate used on this task ($\tau = 0.0005$) is consistent with its higher-variance dynamics and helps stabilize learning. Hopper-v4 shows a different profile: MEOW often improves faster early, whereas TDC-λ catches up and reaches a competitive asymptote, while SAC and DDPG are generally less robust. On Ant-v4 and Walker2d-v4, TD3 tends to be more aggressive in the early phase, but TDC-λ overtakes later and maintains higher final returns. Overall, the curves suggest that TDC-λ trades some early aggressiveness for stronger asymptotic return and stability, consistent with the conservative target selection imposed by λ.

Figure 2 contrasts the two inference modes of TDC-λ. Ant-v4 favors the deterministic actor, whereas Walker2d-v4 benefits from the stochastic tanh-Gaussian head; Hopper-v4 ends with comparable asymptotes, with a mild early-phase advantage for the stochastic head. The most notable result is policy-mode invariance on HalfCheetah-v4 and Humanoid-v4: both actors converge to nearly identical final returns, despite these tasks often being associated with opposite preferences in the literature. This pattern points to the primary role of our critics' LCB-shaped distributional target (per-transition $\mu - \lambda\sigma$) in stabilizing learning, making performance far less sensitive to whether the actor samples or acts deterministically. Implementation-wise, both modes share a single training pipeline. The stochastic head is a tanh-Gaussian policy trained with automatic temperature tuning while evaluation uses mean-action inference for both modes to remove test-time sampling noise. Furthermore,

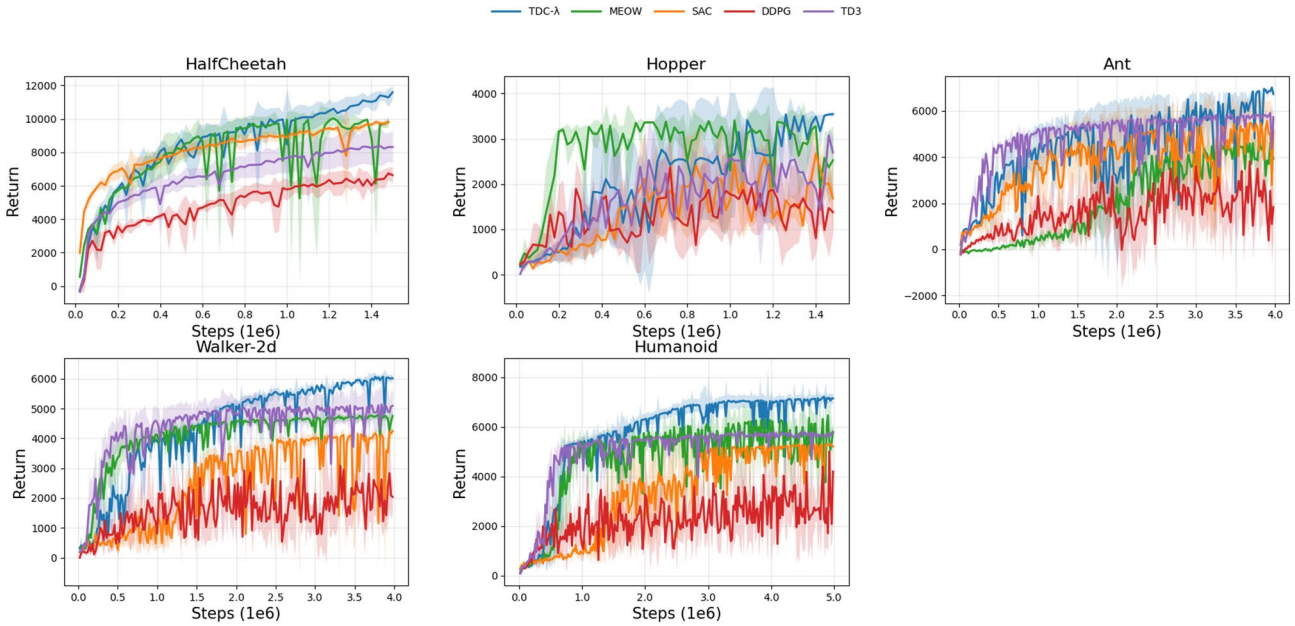

Fig. 1. Performance comparison of the algorithms (TDC-λ, SAC, MEOW, DDPG, TD3).

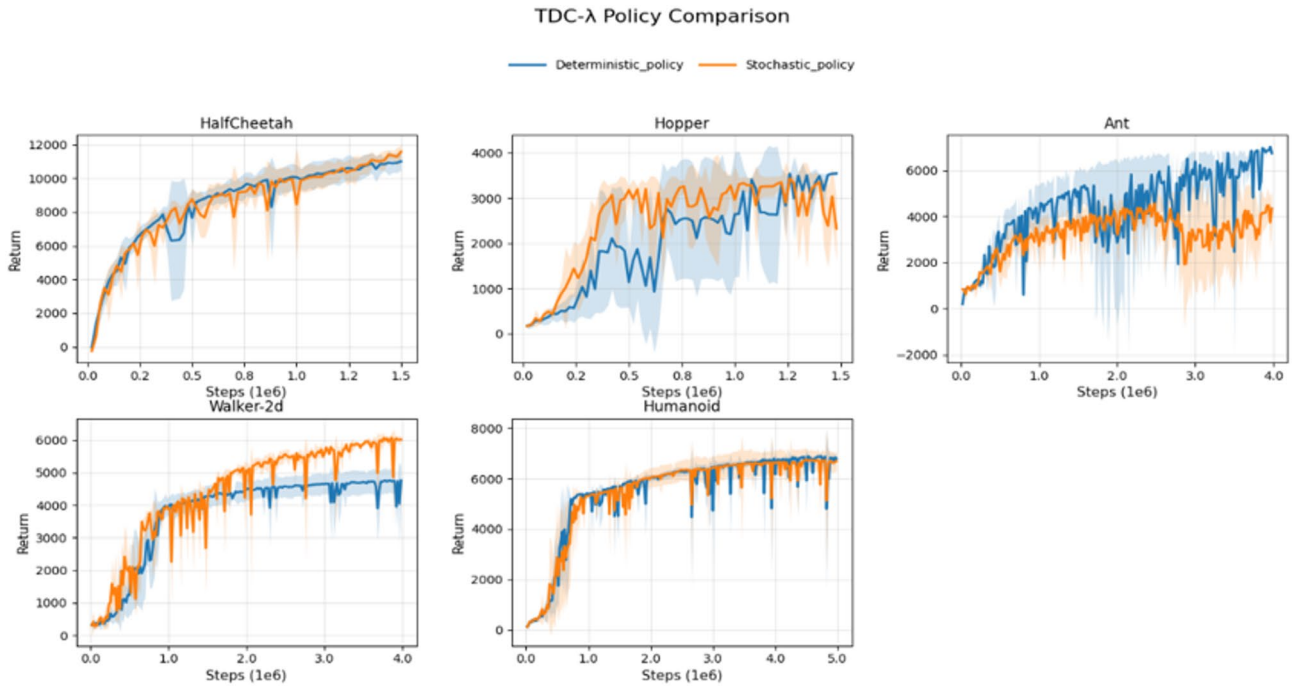

Fig. 2. Performance comparison between TDC-λ with deterministic policy and TDC-λ with stochastic policy.

we sweep $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}$ on Humanoid-v4 (high-dimensional, long-horizon) and HalfCheetah-v4 (lower-dimensional, more stochastic) to probe the impact of our lower-confidence-bound target ($\mu - \lambda\sigma$).

Figure 3 shows that larger λ (≥ 0.75) consistently tightens confidence bands; on Humanoid it also yields the best asymptotic returns, indicating that stronger penalization of high-dispersion critics stabilizes learning in high-variance regimes. Accordingly, we use high λ for higher-variance tasks and low λ for more stable tasks; the final per-environment λ values used in our main experiments are reported in Table 1.

Finally, the box-plot analysis in Fig. 4 highlights robustness after training. Each panel summarizes evaluation returns computed over 100 test episodes per run. Across all five MuJoCo tasks, TDC-λ concentrates mass at higher returns with consistently tighter interquartile ranges, indicating both stronger central performance and greater reliability. On HalfCheetah-v4 and Walker2d-v4, TDC-λ attains the top medians with compact

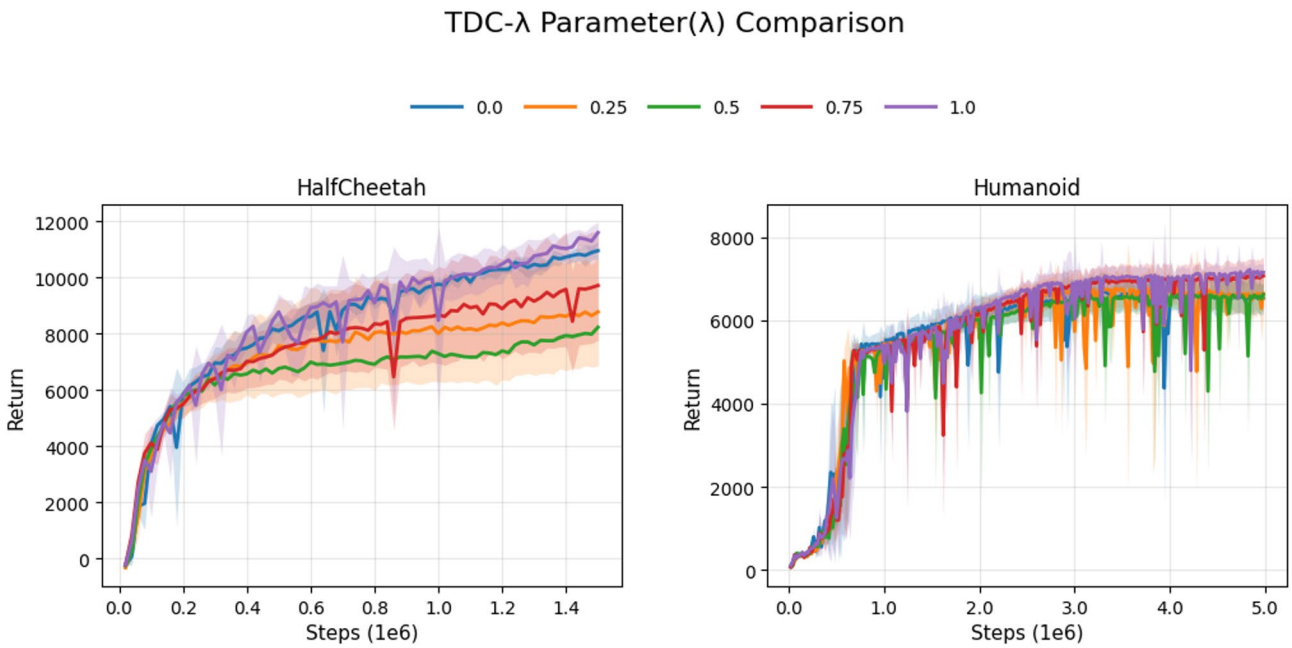

Fig. 3. Comparison of different λ values in TDC-λ algorithm.

Fig. 4. Evaluation of the after train results.

spread; SAC/MEOW form a mid-tier below it, while TD3 is competitive but exhibits a visibly wider dispersion and DDPG lags with substantially lower returns. On Hopper-v4, TDC-λ achieves the tightest high-return distribution; MEOW (and to a lesser extent TD3) can reach similar upper values but shows a pronounced lower tail, whereas SAC and DDPG remain clearly lower overall. On Ant-v4, TDC-λ's box is high and tight; TD3 is relatively close but more variable, and MEOW/SAC exhibit broader, downward-skewed distributions, with DDPG showing the largest low-return tail. On Humanoid-v4, TDC-λ again leads with narrow dispersion; TD3 and SAC are comparatively stable but lower, and MEOW occasionally reaches comparable highs only with much larger variance. Overall, the box plots mirror the learning curves: TDC-λ not only lifts the median but also compresses variability, consistent with its λ-LCB target $(\mu\lambda\sigma)$ mitigating over-optimistic targets and reducing rare catastrophic evaluations.

Environment	Simulator	State dimension	Action dimension
HalfCheetah	PyBullet	26	6
Hopper	PyBullet	15	3
Ant	PyBullet	28	8
Humanoid	PyBullet	44	17
AllegroHand	Isaac	72	16
FrankaCabinet	Isaac	23	9

Table 2. State/action dimensionality of the additional scalability benchmarks (PyBullet + Isaac).

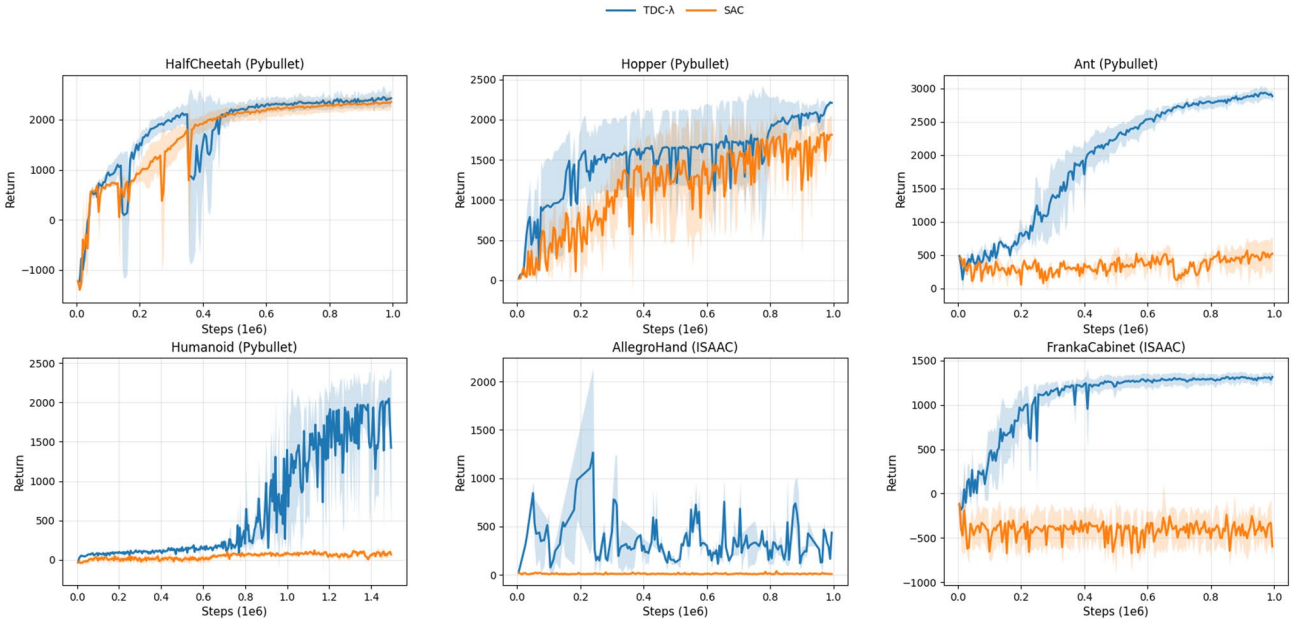

Fig. 5. Scalability evaluation on PyBullet and Isaac environments. Learning curves (average return vs. environment steps) comparing TDC-λ and SAC on four PyBullet locomotion tasks and two Isaac manipulation tasks. Shaded regions indicate variability across independent runs.

Furthermore, beyond final returns, we report training-time stability diagnostics that directly quantify the variability of the bootstrap signal. We measure the minibatch variance of the scalarized $TD\ error\ \delta = \overline{y} - \overline{q}_i$ where $\overline{y}$ is the mean over quantiles of the distributional target and $\overline{q}_i$ is the mean over quantiles of the critic output. On Hopper-v4, $\lambda = 1$ substantially reduces TD-error variability and extreme spikes relative to $\lambda = 0$, while the target policy-smoothing noise statistics remain essentially identical (Supplementary Fig. S1). These diagnostics provide additional evidence that the $\mu - \lambda\sigma$ target selection improves stability by reducing over-optimistic target propagation.

In Table 2 we use observation/action dimensionality and contact complexity as practical proxies for task scale, while also testing cross-simulator robustness.

To evaluate whether the proposed method scales beyond MuJoCo and is not tied to a single physics engine, we conducted additional experiments on two alternative simulators: PyBullet and NVIDIA Isaac (Isaac Lab). We considered four PyBullet locomotion environments (HalfCheetah, Hopper, Ant, Humanoid) and two Isaac manipulation environments (AllegroHand and FrankaCabinet), spanning a broad range of observation/action dimensionalities and contact complexity (Table 2). Figure 5 reports learning curves comparing TDC-λ against SAC. Across all tasks and both simulators, TDC-λ exhibits stable learning and consistently achieves higher (or comparable) returns than SAC; the advantage is particularly clear in contact-rich and higher-dimensional settings (e.g., Ant and Humanoid) and in the FrankaCabinet manipulation task. These results support the scalability of the proposed risk-sensitive distributional target selection across different simulators and control regimes.

Discussion

This study proposed TDC λ, a risk-aware extension of TD3 that combines twin distributional critics with a per-transition lower-confidence-bound (LCB) rule for target construction, while allowing both deterministic and stochastic actors to be trained within a single off-policy pipeline. Rather than modifying the actor objective or introducing additional regularization terms, TDC λ intervenes at the level of the Bellman target: for each

transition, it scores the two distributional critics using $\mu - \lambda\sigma$ and propagates the quantiles of the safer critic. This mechanism biases learning away from over-optimistic critics precisely in high-variance regimes, without relying on task-specific heuristics or architectural complexity beyond replacing scalar Q-functions with quantile heads.

We benchmarked TDC λ against SAC and a recent energy-based flow method (MEOW/EBFlow), which represent strong and widely adopted MaxEnt baselines and TD3 and DDPG as well. Conceptually, SAC alternates critic evaluation and policy improvement under an entropy-regularized objective, whereas EBFlow unifies evaluation and improvement for stochastic policies via normalizing flows. TDC λ follows a different design philosophy: it retains the simple TD3-style off-policy loop but shapes the TD target itself through distributional critics and a tunable risk parameter. The unified training stack exposes a tanh-Gaussian policy during learning while defaulting to deterministic mean actions at evaluation, a regime that has often proved competitive in continuous control. Across five MuJoCo tasks, this configuration yielded higher or statistically comparable asymptotic returns to both SAC, MEOW, TD3 and DDPG, while consistently reducing run-to-run variability. To place these results in a broader context, we contrasted them with the step-aligned MuJoCo learning curves reported by Chao et al.[7] or DDPG, and vanilla TD3 on Hopper-v4, HalfCheetah-v4, Walker2d-v4, Ant-v4, and Humanoid-v4. In that unified evaluation, these canonical on- and off-policy baselines typically plateau at lower returns and exhibit wider performance bands than modern MaxEnt approaches. In our study, we run MaxEnt baselines (SAC and MEOW) and DDPG/TD3. Relative to these established methods, TDC λ consistently occupies the higher-performing regime and exhibits tighter dispersion, suggesting that a significant portion of the empirical gains associated with MaxEnt training can be recovered by risk-aware target shaping within a TD3 framework. Our earlier AdvB SAC and AdvB TD3 algorithms targeted variance at the decision level, using an advisory ensemble to choose among candidate actions scored by a shared critic[19]. TDC λ acts at a different point in the pipeline: two independent distributional critics predict return quantiles, and for each transition the algorithm selects one critic via the LCB score $\mu - \lambda\sigma$ before constructing the quantile target. Empirically, TDC λ matched or surpassed AdvB SAC/TD3 across HalfCheetah-v4, Walker2d-v4, Ant-v4, and Humanoid-v4, achieving higher or comparable asymptotic returns with narrower performance distributions, and remained competitive on Hopper-v4. These results indicate that stabilizing the target can be at least as effective as more elaborate action-selection schemes, while keeping the actor architecture and training loop simple.

Policy-mode and λ ablations further clarify the role of risk sensitivity in TDC λ. Some environments modestly favored deterministic evaluation, whereas others benefited slightly from the stochastic tanh-Gaussian actor, but on the more challenging tasks both modes converged to similar performance. This supports the view that the main advantage of TDC λ lies in stabilizing learning rather than enforcing a particular exploration strategy. Varying λ showed that larger values are especially beneficial in high-dimensional settings with noisy dynamics and dispersed critics, while $\lambda \approx 0$ behaves similarly to a distributional analogue of clipped double TD3 when critic uncertainty is benign. Together, these findings support the interpretation of TDC λ as a practical methodological bridge between deterministic TD3 and stochastic MaxEnt training, and motivate future work on adaptive λ schedules and extensions beyond the MuJoCo suite.

TDC-λ introduces a single additional risk hyperparameter λ. While λ is low-dimensional and can be tuned with a modest search budget, performance can be sensitive in some tasks; therefore, automated selection (as used in our experiments) or adaptive λ schedules are promising directions for improving applicability without per-task tuning.

Finally, while the λ-LCB rule has a clear statistical interpretation as a conservative bound, we do not claim a formal convergence guarantee for deep off-policy learning with nonlinear function approximation; instead, we evaluate its effect empirically via performance and additional stability diagnostics (TD-target variability and TD-error variability) and treat λ as a tunable conservatism parameter.

Conclusion

We introduced TDC λ, a twin-critic distributional variant of TD3 that constructs risk-sensitive TD targets by scoring each critic with a per-transition lower confidence bound $\mu - \lambda\sigma$ and propagating the quantiles of the safer critic. This drop-in modification replaces scalar Q-values with quantile critics and exposes a single off-policy pipeline that supports both deterministic and stochastic actors while keeping test-time evaluation deterministic by default. Across five standard MuJoCo benchmarks, TDC λ achieved higher or statistically comparable asymptotic returns to SAC and an energy-based flow baseline, and produced markedly tighter performance distributions over independent runs. Actor-mode ablations indicated that both deterministic and stochastic policies benefit from the same LCB-based target-shaping mechanism, while λ sweeps suggested that the risk parameter largely controls stability and variance: larger λ values tend to be most advantageous when critic dispersion is high, whereas λ close to zero often suffices when the critics are already well behaved.

These results suggest that much of the robustness commonly attributed to sophisticated MaxEnt architectures can instead be realized by shaping the learning target with a simple distributional LCB rule, without abandoning the practical advantages of TD3-style off-policy training. More broadly, they highlight distributional and risk-aware target construction as a promising design axis for reinforcement learning algorithms that aim to combine high asymptotic performance with improved stability across tasks.

Data availability

https://github.com/tkaraca/Risk-sensitive-twin-distributional-critics-with-a-lambda-lower-confidence-bound.

Received: 25 November 2025; Accepted: 27 January 2026
Published online: 30 January 2026

References

1. Lillicrap, T. P. et al. Continuous control with deep reinforcement learning. Preprint at https://doi.org/10.48550/ARXIV.1509.02971 (2015).
2. Fujimoto, S., van Hoof, H. & Meger, D. Addressing function approximation error in actor-critic methods. Preprint at https://doi.org/10.48550/arXiv.1802.09477 (2018).
3. Hasselt, H. V. Double Q-learning (2010).
4. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
5. Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Preprint at https://doi.org/10.48550/arXiv.1801.01290 (2018).
6. Haarnoja, T., Tang, H., Abbeel, P. & Levine, S. Reinforcement learning with deep energy-based policies. Preprint at https://doi.org/10.48550/arXiv.1702.08165 (2017).
7. Chao, C.-H. et al. Maximum entropy reinforcement learning via energy-based normalizing flow.
8. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density estimation using real NVP. Preprint at https://doi.org/10.48550/arXiv.1605.08803 (2017).
9. Kingma, D. P. & Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. Preprint at https://doi.org/10.48550/arXiv.1807.03039 (2018).
10. Papamakarios, G., Murray, I. & Pavlakou, T. Masked autoregressive flow for density estimation.
11. Durkan, C., Bekasov, A., Murray, I. & Papamakarios, G. Neural spline flows. Preprint at https://doi.org/10.48550/arXiv.1906.04032 (2019).
12. Nachum, O., Norouzi, M., Xu, K. & Schuurmans, D. Trust-PCL: An off-policy trust region method for continuous control. Preprint at https://doi.org/10.48550/arXiv.1707.01891 (2018).
13. Gu, S., Lillicrap, T., Ghahramani, Z., Turner, R. E. & Levine, S. Q-Prop: Sample-efficient policy gradient with an off-policy critic. Preprint at https://doi.org/10.48550/arXiv.1611.02247 (2017).
14. Meng, W., Zheng, Q., Shi, Y. & Pan, G. An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 2223–2235 (2022).
15. Wu, J., Wu, Q. M. J., Chen, S., Pourpanah, F. & Huang, D. A-TD3: An adaptive asynchronous twin delayed deep deterministic for continuous action spaces. *IEEE Access* **10**, 128077–128089 (2022).
16. Chowdhury, M. A., Al-Wahaibi, S. S. S. & Lu, Q. Entropy-maximizing TD3-based reinforcement learning for adaptive PID control of dynamical systems. *Comput. Chem. Eng.* **178**, 108393 (2023).
17. Xu, T., Meng, Z., Lu, W. & Tong, Z. End-to-end autonomous driving decision method based on improved TD3 algorithm in complex scenarios. *Sensors* **24**, 4962 (2024).
18. Luo, B., Wu, Z., Zhou, F. & Wang, B.-C. Human-in-the-loop reinforcement learning in continuous-action space. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 15735–15744 (2024).
19. Osman, O., Karaca, T. K., Kavus, B. Y. & Tulum, G. AdvB-TD3: A novel decision-making framework for complex continuous control tasks. *IEEE Access* **13**, 181675–181685 (2025).

Author contributions

Conceptualization, T.K.K. and O.O.; methodology, O.O., T.K.K and B.Y.K.; validation, M.A.K., and B.Y.K.; formal analysis, T.K.K. and G.T.; data curation, B.Y.K, T.K.K. and G.T.; writing—original draft preparation, B.Y.K and T.K.K.; writing—review and editing, M.A.K, G.T. and B.Y.K and T.K.K.; visualization, B.Y.K and G.T.; supervision, O.O. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-37910-3.

Correspondence and requests for materials should be addressed to T.K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.